This is a video transcript of Module 5 of the Online Open Course in Species Distribution Modelling. To access the full suite of videos click here.

Online Open Course - Species Distribution Modelling

Powered by the Biodiversity and Climate Change Virtual Laboratory

---

Module 5 Presence only models

In the previous modules of this online open course in species distribution modelling, we have learnt a great deal about the kind of data that you need to build a model, and we had a quick look at the different kinds of algorithms that you can use. So, now it is time to explain the algorithms in more detail. In this module we will look at models that only use presence data. I will explain two different geographic models and then move to Bioclim, which is the most well known profile model. And lastly I will explain Maxent, a popular machine learning model.

An easy way to start exploring the distribution of your species of interest is by mapping the points where the species has been recorded and building a geographic model. These geographic models only use presence data, and simply draw a shape around them. They are mostly useful to obtain a quick understanding of the area where a species occurs.

Geographic models

Let's start with geographic models. As I mentioned in module 4, geographic models are an easy way to start exploring the distribution of your species of interest. These models only use presence data, and simply draw a shape around them. Because they don't use any environmental data, they are mostly useful to obtain a quick idea of where a species occurs.

Two examples of geographic models are Circles and Convex Hull. The Circles model predicts that a species can be present within a circle of some given radius around any observed occurrence point. The Convex Hull model draws a spatial convex hull around the set of occurrence points. This is the smallest polygon that you can draw around the points of occurrences enclosing all points, but where all the angles of the polygon boundary are outwardly convex. For any two points, a line between them has to fall completely within the convex hull. The model predicts that a species can be present in any location within the convex hull, and will be absent outside the convex hull.

The Convex Hull model is, for example, used by the International Union for Conservation of Nature (IUCN) to estimate the extent of occurrence for species. IUCN considers a species to be

critically endangered if the size of the convex hull is less than 100 km². In Australia, the northern hairy-nosed wombat is an example of a critically endangered species. Its range is restricted to a 3 km^2 extent within the Epping Forest National Park in Queensland.

These geographic models are easy to understand and interpret, but not very often used because they don't take into account of the environmental conditions that regulate or influence a species.

Profile models: Bioclim

The next type of models are profile models, which are the most basic actual species distribution models. They profile the environments where a species can be found. The best known profile method is Bioclim, which is considered as the first species distribution model. Development of this model started in Australia in the mid-1960s.

Bioclim is a relatively simple and easy to understand model. While you would usually take into account more than two environmental conditions as predictors, I visualize the Bioclim concept here with a two-dimensional graph with two environmental variables, and a set of occurrence points. Bioclim constructs the suitable environmental space as a bounding box around the minimum and maximum values of the environmental variables for all occurrences. The green area within the bounding box represents the suitable environmental conditions for a species to survive. To avoid the over-predictive effect of outliers, the resulting envelope can be reduced to user-specified percentiles, such as the inner 95%. If you include more than two environmental variables in the model, this would result in a multi-dimensional bounding box.

To predict the probability of species occurrences in any given location, Bioclim compares the values of the environmental variables at that location to the percentile distribution of the values from known occurrence locations. The 50^{th} percentile refers to the median, which divides the data exactly in half. The closer the value of the environmental variable at the unknown location is to the 50^{th} percentile, the more suitable the location is for a species to occur there, and thus the higher the probability of occurrence. So, at the 50^{th} percentile, the probability is 1, and the probability of the 10^{th} percentile is the same as the probability of the 90^{th} percentile. The Bioclim model combines the scores for each environmental variable into an overall probability of occurrence for each location with equal weights for all environmental variables.

Although Bioclim has been widely used for species distribution modelling, it has a few limitations. The method can only handle continuous environmental variables, and thus not take into account categorical variables such as soil type. And it does not take into account interactions between variables. With the development of more complex algorithms, this algorithm is therefore less and less used for species distribution modelling.

Maxent

One of these more complex algorithms is Maxent, which stands for maximum entropy modelling. Maxent is an algorithm that only uses presence data, and it compares the locations of where a

species has been found to all the environments that are available in the study region. It defines these available environments by sampling a large number of points throughout the study area. These points are called background points. Because background points can include locations where the species is known to occur, background points are not the same as pseudo-absence points. Background points define the available environment.

Maxent has two main components. The first is entropy, this means that the model is calibrated to find the distribution that is most spread out, or closest to uniform throughout the study region. The second is constraint, which are the rules that constrain the predicted distribution. These rules are based on the values of the environmental variables of the locations where the species has been observed. One of these constraints can be that the mean of each environmental variable at an unknown locations must be close to the mean of those variables in known occurrence locations.

I will illustrate this with a simple example. This grid represents a landscape, in which in some cells we have observed a bird species to be present. For the other cells, we do not know whether the species is present or absent. One of the environmental variables that we put into the model is the mean annual temperature and we have a value of this environmental variable for each cell. Maxent calculates the mean of all cells in which the species has been recorded, which in this case is 20 degree celsius. The first constraint is that the mean annual temperature of all cells in which the species is predicted to be present must be 20 degrees. The second constraint is that the sum of the predicted probabilities across all cells in the landscape needs to be equal to 1. One of the possible distributions that meets both of these constraints is to give each cell where the species has been observed a probability of 1/4, and the other cells a probability of 0. The mean temperature of the cells in which the species is predicted to be present is 20 degrees, meeting constraint 1; and the probabilities also sum up to 1, which means constraint 2 is met. Another option is to give 12 cells a probability of 1/12, which also sums up to 1, and the mean temperature across these 12 cells is also 20 degrees. The 4 cells with a mean temperature of 40 degrees get a probability of 0. This distribution also meets both constraints. A third option is to give more weight to the cells where the species has been observed compared to the other cells. This option also meets both constraints. This is just an illustration of three options, but there are more potential distributions which meet the constraint that the mean annual temperature of the cells with a predicted presence is 20 degrees. Maxent then queries which of these potential distributions is the most uniform, which in this example is option 2. While this example only included one environmental variable, Maxent of course uses a suite of variables to model the distribution of a species, and these are called features.

Maxent considers six types of features, and each of these types allows a different possible shape of the response curves, and has different implications for the constraints. The first feature type is linear, which refers to a continuous variable. The constraint for linear features is as I explained in the example: the mean of each environmental variable at an unknown location must be close to the mean of those variables in known occurrence locations. The second type is

quadratic, which is the square of a continuous variable and accounts for the species' tolerance for variation from its optimal conditions. The third feature type is product, which allows for interactions between variables. Then there is the threshold feature type, which converts a continuous response in a binary response. The hinge feature type is a combination of the linear and threshold type. The last feature type is categorical, which refers to categorical variables with different classes of the response. As a default Maxent uses all feature types, but you can choose to build simpler models by only using a few of these.

To calculate the potential distribution of a species, Maxent first calculates two probability densities. For all presence points, the probability density describes the relative likelihood of all environmental variables in the model over the range of the points. So, in this case, across all the presence points, the values for temperature and rainfall under the peak in the graph on the right were the most common. The same is done across the entire study region, based on the background points. So, the probability density of the background points characterizes the available environment within the study region, whereas the probability density of the presence points characterizes the environment of where a species has been found. Maxent then calculates the ratio between these two probability densities, which gives the relative environmental suitability for presence of a species, for each point in the study area.

Maxent chooses the distribution that maximizes the similarity between the environmental characteristics of the total environment and those of the locations where the species is known to be present. This is known as the raw output of Maxent. For easier interpretation of the results, and to provide an estimate of the probability that a species is present in a given location, Maxent performs a logistic transformation of the raw output. The logistic output takes into account the prevalence of a species, which refers to the proportion of occupied locations. However, the exact prevalence cannot be derived from presence-only data. The value for prevalence can be defined by the user, but Maxent uses a default of 0.5. We advise to take care with this default as this is not appropriate for rare species.

An important aspect of Maxent is regularization, which reduces overfitting of the model. Regularization is done in two ways: first, Maxent relaxes the constraints, and instead of fitting the model around the exact means of the environmental variables, it takes into account the confidence intervals around the means. It thus prevents the model from being fitted to closely around the input data. Secondly, the model uses a penalty for complexity, which means that it excludes feature types that don't add a significant improvement to the model.

I hope this module has given you a better understanding of how these presence-only models work. I am looking forward to seeing you back in the next module in which we discuss how statistical regression models work.

Attribution

Please cite this video as follows:

Huijbers CM, Richmond SJ, Low-Choy SJ, Laffan SW, Hallgren W, Holewa H (2016) SDM Online Open Course, module 5: presence only models. Biodiversity and Climate Change Virtual Laboratory, http://www.bccvl.org.au/training/. DDMMYY of access.

Acknowledgements

Lead partners: Griffith University, James Cook University

Thanks to: University of New South Wales, Macquarie University, University of Canberra

Funded by: National eResearch Tools and Resources Project (NeCTAR)

The BCCVL is supported by the National eResearch Tools and Resources Project (NeCTAR), an initiative of the Commonwealth being conducted as part of the Super Science Initiative and financed from the Education Investment Fund, Department of Education. The University of Melbourne is the lead agent for the delivery of the NeCTAR project and Griffith University is the sub-contractor.